

**Techniques of Water-Resources Investigations of the United States Geological Survey**

**Book 4, Hydrologic Analysis and Interpretation**

**Chapter A3**

# **Statistical Methods in Water Resources**

**By D.R. Helsel and R.M. Hirsch**

# Table of Contents

---

Preface	xi
<b>Chapter 1 Summarizing Data</b>	<b>1</b>
1.1 Characteristics of Water Resources Data	2
1.2 Measures of Location	3
1.2.1 Classical Measure -- the Mean	3
1.2.2 Resistant Measure -- the Median	5
1.2.3 Other Measures of Location	6
1.3 Measures of Spread	7
1.3.1 Classical Measures	7
1.3.2 Resistant Measures	8
1.4 Measures of Skewness	9
1.4.1 Classical Measure of Skewness	9
1.4.2 Resistant Measure of Skewness	10
1.5 Other Resistant Measures	10
1.6 Outliers	11
1.7 Transformations	12
1.7.1 The Ladder of Powers	12
<b>Chapter 2 Graphical Data Analysis</b>	<b>17</b>
2.1 Graphical Analysis of Single Data Sets	19
2.1.1 Histograms	19
2.1.2 Stem and Leaf Diagrams	20
2.1.3 Quantile Plots	22
2.1.4 Boxplots	24
2.1.5 Probability Plots	27
2.2 Graphical Comparisons of Two or More Data Sets	35
2.2.1 Histograms	35
2.2.2 Dot and Line Plots of Means, Standard Deviations	35
2.2.3 Boxplots	38
2.2.4 Probability Plots	40
2.2.5 Q-Q Plots	41
2.3 Scatterplots and Enhancements	45

2.3.1	Evaluating Linearity	45
2.3.2	Evaluating Differences in Location on a Scatterplot	47
2.3.3	Evaluating Differences in Spread	50
2.4	Graphs for Multivariate Data	51
2.4.1	Profile Plots	51
2.4.2	Star Plots	53
2.4.3	Trilinear Diagrams	56
2.4.4	Plots of Principal Components	58
2.4.5	Other Multivariate Plots	59
<b>Chapter 3</b>	<b>Describing Uncertainty</b>	<b>65</b>
3.1	Definition of Interval Estimates	66
3.2	Interpretation of Interval Estimates	67
3.3	Confidence Intervals For The Median	70
3.3.1	Nonparametric Interval Estimate For The Median	70
3.3.2	Parametric Interval Estimate For The Median	73
3.4	Confidence Intervals For The Mean	74
3.4.1	Symmetric Confidence Interval For The Mean	75
3.4.2	Asymmetric Confidence Interval For The Mean	76
3.5	Nonparametric Prediction Intervals	76
3.5.1	Two-Sided Nonparametric Prediction Interval	77
3.5.2	One-Sided Nonparametric Prediction Interval	78
3.6	Parametric Prediction Intervals	79
3.6.1	Symmetric Prediction Interval	80
3.6.2	Asymmetric Prediction Intervals	80
3.7	Confidence Intervals For Percentiles (Tolerance Intervals)	81
3.7.1	Nonparametric Confidence Intervals For Percentiles	83
3.7.2	Nonparametric Tests For Percentiles	84
3.7.3	Parametric Confidence Intervals For Percentiles	88
3.7.4	Parametric Tests For Percentiles	89
3.8	Other Uses For Confidence Intervals	90
3.8.1	Implications of Non-Normality For Detection of Outliers	90
3.8.2	Implications of Non-Normality For Quality Control	91
3.8.3	Implications of Non-Normality For Sampling Design	93
<b>Chapter 4</b>	<b>Hypothesis Tests</b>	<b>97</b>
4.1	Classification of Hypothesis Tests	99
4.1.1	Classification Based on Measurement Scales	99
4.1.2	Classification Based on the Data Distribution	100

4.2	Structure of Hypothesis Tests	101
4.2.1	Choose the Appropriate Test	101
4.2.2	Establish the Null and Alternate Hypotheses	104
4.2.3	Decide on an Acceptable Error Rate $\alpha$	106
4.2.4	Compute the Test Statistic from the Data	108
4.2.5	Compute the p-value	108
4.2.6	Make the Decision to Reject $H_0$ or Not	108
4.3	The Rank-Sum Test as an Example of Hypothesis Testing	109
4.4	Tests for Normality	113
<b>Chapter 5</b>	<b>Differences Between Two Independent Groups</b>	<b>117</b>
5.1	The Rank-Sum Test	118
5.1.1	Null and Alternate Hypotheses	118
5.1.2	Computation of the Exact Test	119
5.1.3	The Large-Sample Approximation	121
5.1.4	The Rank Transform Approximation	123
5.2	The t-Test	124
5.2.1	Assumptions of the Test	124
5.2.2	Computation of the t-Test	125
5.2.3	Modification for Unequal Variances	125
5.2.4	Consequences of Violating the t-Test's Assumptions	127
5.3	Graphical Presentation of Results	128
5.3.1	Side-by-Side Boxplots	128
5.3.2	Q-Q Plots	129
5.4	Estimating the Magnitude of Differences Between Two Groups	131
5.4.1	The Hodges-Lehmann Estimator	131
5.4.2	Confidence interval for $\hat{D}$	132
5.4.3	Difference Between Mean Values	134
5.4.4	Confidence interval for $\bar{x} - \bar{y}$	134
<b>Chapter 6</b>	<b>Matched-Pair Tests</b>	<b>137</b>
6.1	The Sign Test	138
6.1.1	Null and Alternate Hypotheses	138
6.1.2	Computation of the Exact Test	138
6.1.3	The Large-Sample Approximation	141
6.2	The Signed-Rank Test	142
6.2.1	Null and Alternate Hypotheses	142
6.2.2	Computation of the Exact Test	143
6.2.3	The Large-Sample Approximation	145
6.2.4	The Rank Transform Approximation	147

6.3	The Paired t-Test	147
6.3.1	Assumptions of the Test	147
6.3.2	Computation of the Paired t-Test	148
6.4	Consequences of Violating Test Assumptions	149
6.4.1	Assumption of Normality (t-Test)	149
6.4.2	Assumption of Symmetry (Signed-Rank Test)	150
6.5	Graphical Presentation of Results	150
6.5.1	Boxplots	151
6.5.2	Scatterplots With X=Y Line	151
6.6	Estimating the Magnitude of Differences Between Two Groups	153
6.6.1	The Median Difference (Sign Test)	153
6.6.2	The Hodges-Lehmann Estimator (Signed-Rank Test)	153
6.6.3	Mean Difference (t-Test)	155
<b>Chapter 7</b>	<b>Comparing Several Independent Groups</b>	<b>157</b>
7.1	Tests for Differences Due to One Factor	159
7.1.1	The Kruskal-Wallis Test	159
7.1.2	Analysis of Variance (One Factor)	164
7.2	Tests For The Effects of More Than One Factor	169
7.2.1	Nonparametric Multi-Factor Tests	170
7.2.2	Multi-Factor Analysis of Variance -- Factorial ANOVA	170
7.3	Blocking -- The Extension of Matched-Pair Tests	181
7.3.1	Median Polish	182
7.3.2	The Friedman Test	187
7.3.3	Median Aligned-Ranks ANOVA	191
7.3.4	Parametric Two-Factor ANOVA Without Replication	193
7.4	Multiple Comparison Tests	195
7.4.1	Parametric Multiple Comparisons	196
7.4.2	Nonparametric Multiple Comparisons	201
7.5	Presentation of Results	202
7.5.1	Graphical Comparisons of Several Independent Groups	202
7.5.2	Presentation of Multiple Comparison Tests	205
<b>Chapter 8</b>	<b>Correlation</b>	<b>209</b>
8.1	Characteristics of Correlation Coefficients	210
8.1.1	Monotonic Versus Linear Correlation	210
8.2	Kendall's Tau	212
8.2.1	Computation	212
8.2.2	Large Sample Approximation	213
8.2.3	Correction for Ties	215

8.3	Spearman's Rho	217
8.4	Pearson's $r$	218
<b>Chapter 9 Simple Linear Regression</b>		<b>221</b>
9.1	The Linear Regression Model	222
9.1.1	Assumptions of Linear Regression	224
9.2	Computations	225
9.2.1	Properties of Least Squares Solutions	227
9.3	Building a Good Regression Model	228
9.4	Hypothesis Testing in Regression	237
9.4.1	Test for Whether the Slope Differs From Zero	237
9.4.2	Test for Whether the Intercept Differs from Zero	238
9.4.3	Confidence Intervals on Parameters	239
9.4.4	Confidence Intervals for the Mean Response	240
9.4.5	Prediction Intervals for Individual Estimates of $y$	241
9.5	Regression Diagnostics	244
9.5.1	Measures of Outliers in the $x$ Direction	246
9.5.2	Measures of Outliers in the $y$ Direction	246
9.5.3	Measures of Influence	248
9.5.4	Measures of Serial Correlation	250
9.6	Transformations of the Response ( $y$ ) Variable	253
9.6.1	To Transform or Not to Transform?	253
9.6.2	Consequences of Transformation of $y$	253
9.6.3	Computing Predictions of Mass (Load)	255
9.6.4	An Example	257
9.7	Summary Guide to a Good SLR Model	261
<b>Chapter 10 Alternative Methods to Regression</b>		<b>265</b>
10.1	Kendall-Theil Robust Line	266
10.1.1	Computation Of The Line	266
10.1.2	Properties Of The Estimator	267
10.1.3	Test Of Significance	272
10.1.4	Confidence Interval For Theil Slope	273
10.2	Alternative Parametric Linear Equations	274
10.2.1	OLS Of $X$ On $Y$	275
10.2.2	Line of Organic Correlation	276
10.2.3	Least Normal Squares	278
10.2.4	Summary Of The Applicability Of OLS, LOC And LNS	280
10.3	Weighted Least Squares	280
10.4	Iteratively Weighted Least Squares	283

10.5 Smoothing	285
10.5.1 Moving median smooths	285
10.5.2 LOWESS	287
10.5.3 Polar smoothing	291
<b>Chapter 11 Multiple Regression</b>	<b>295</b>
11.1 Why Use MLR?	296
11.2 MLR Model	296
11.3 Hypothesis Tests for Multiple Regression	297
11.3.1 Nested F Tests	297
11.3.2 Overall F Test	298
11.3.3 Partial F Tests	298
11.4 Confidence Intervals	299
11.4.1 Variance-Covariance Matrix	299
11.4.2 Confidence Intervals for Slope Coefficients	300
11.4.3 Confidence Intervals for the Mean Response	300
11.4.4 Prediction Intervals for an Individual $y$	300
11.5 Regression Diagnostics	300
11.5.1 Partial Residual Plots	301
11.5.2 Leverage and Influence	301
11.5.3 Multi-Collinearity	305
11.6 Choosing the Best MLR Model	309
11.6.1 Stepwise Procedures	310
11.6.2 Overall Measures of Quality.	313
11.7 Summary of Model Selection Criteria	315
11.8 Analysis of Covariance	316
11.8.1 Use of One Binary Variable	316
11.8.2 Multiple Binary Variables	318
<b>Chapter 12 Trend Analysis</b>	<b>323</b>
12.1 General Structure of Trend Tests	324
12.1.1 Purpose of Trend Testing	324
12.1.2 Approaches to Trend Testing	325
12.2 Trend Tests With No Exogenous Variable	326
12.2.1 Nonparametric Mann-Kendall Test	326
12.2.2 Parametric Regression of $Y$ on $T$	328
12.2.3 Comparison of Simple Tests for Trend	328
12.3 Accounting for Exogenous Variables	329
12.3.1 Nonparametric Approach	334
12.3.2 Mixed approach	335

12.3.3	Parametric Approach	335
12.3.4	Comparison of Approaches	336
12.4	Dealing With Seasonality	337
12.4.1	The Seasonal Kendall Test	338
12.4.2	Mixture Methods	340
12.4.3	Multiple Regression With Periodic Functions	342
12.4.4	Comparison of Methods	342
12.4.5	Presenting Seasonal Effects	343
12.4.6	Differences Between Seasonal Patterns	344
12.5	Use of Transformations in Trend Studies	346
12.6	Monotonic Trend versus Two Sample (Step) Trend	348
12.7	Applicability of Trend Tests With Censored Data	352
<b>Chapter 13</b>	<b>Methods for Data Below the Reporting Limit</b>	<b>357</b>
13.1	Methods for Estimating Summary Statistics	358
13.1.1	Simple Substitution Methods	358
13.1.2	Distributional Methods	360
13.1.3	Robust Methods	362
13.1.4	Recommendations	362
13.1.5	Multiple Reporting Limits	364
13.2	Methods for Hypothesis Testing	366
13.2.1	Simple Substitution Methods	366
13.2.2	Distributional Test Procedures	367
13.2.3	Nonparametric Tests	367
13.2.4	Hypothesis Testing With Multiple Reporting Limits	369
13.2.5	Recommendations	370
13.3	Methods For Regression With Censored Data	371
13.3.1	Kendall's Robust Line Fit	371
13.3.2	Tobit Regression	371
13.3.3	Logistic Regression	372
13.3.4	Contingency Tables	373
13.3.5	Rank Correlation Coefficients	373
13.3.6	Recommendations	374
<b>Chapter 14</b>	<b>Discrete Relationships</b>	<b>377</b>
14.1	Recording Categorical Data	378
14.2	Contingency Tables (Both Variables Nominal)	378
14.2.1	Performing the Test for Independence	379
14.2.2	Conditions Necessary for the Test	381
14.2.3	Location Of the Differences	382
14.3	Kruskal-Wallis Test for Ordered Categorical Responses	382



14.3.1	Computing the Test	383
14.3.2	Multiple Comparisons	385
14.4	Kendall's Tau for Categorical Data (Both Variables Ordinal)	385
14.4.1	Kendall's $\tau_b$ for Tied Data	385
14.4.2	Test Of Significance for $\tau_b$	388
14.5	Other Methods for Analysis of Categorical Data	390
<b>Chapter 15</b>	<b>Regression for Discrete Responses</b>	<b>393</b>
15.1	Regression For Binary Response Variables.	394
15.1.1	Use of Ordinary Least Squares	394
15.2	Logistic Regression	395
15.2.1	Important Formulae	395
15.2.2	Computation by Maximum Likelihood	396
15.2.3	Hypothesis Tests	397
15.2.4	Amount of Uncertainty Explained, $R^2$	398
15.2.5	Comparing Non-Nested Models	398
15.3	Alternatives to Logistic Regression	402
15.3.1	Discriminant Function Analysis	402
15.3.2	Rank-Sum Test	402
15.4	Logistic Regression for More Than Two Response Categories	403
15.4.1	Ordered Response Categories	403
15.4.2	Nominal Response Categories	405
<b>Chapter 16</b>	<b>Presentation Graphics</b>	<b>409</b>
16.1	The Value of Presentation Graphics	410
16.2	Precision of Graphs	411
16.2.1	Color	412
16.2.2	Shading	413
16.2.3	Volume and Area	416
16.2.4	Angle and Slope	417
16.2.5	Length	420
16.2.6	Position Along Nonaligned Scales	421
16.2.7	Position Along an Aligned Scale	423
16.3	Misleading Graphics To Be Avoided	423
16.3.1	Perspective	423
16.3.2	Graphs With Numbers	426
16.3.3	Hidden Scale Breaks	427
16.3.4	Overlapping Histograms	428
<b>References</b>		<b>433</b>

Appendix A	Construction of Boxplots	451
Appendix B	Tables	456
Appendix C	Data Sets	468
Appendix D	Answers to Exercises	469
Index		500



## Preface

This book began as class notes for a course we teach on applied statistical methods to hydrologists of the Water Resources Division, U. S. Geological Survey (USGS). It reflects our attempts to teach statistical methods which are appropriate for analysis of water resources data. As interest in this course has grown outside of the USGS, incentive grew to develop the material into a textbook. The topics covered are those we feel are of greatest usefulness to the practicing water resources scientist. Yet all topics can be directly applied to many other types of environmental data.

This book is not a stand-alone text on statistics, or a text on statistical hydrology. For example, in addition to this material we use a textbook on introductory statistics in the USGS training course. As a consequence, discussions of topics such as probability theory required in a general statistics textbook will not be found here. Derivations of most equations are not presented. Important tables included in all general statistics texts, such as quantiles of the normal distribution, are not found here. Neither are details of how statistical distributions should be fitted to flood data -- these are adequately covered in numerous books on statistical hydrology.

We have instead chosen to emphasize topics not always found in introductory statistics textbooks, and often not adequately covered in statistical textbooks for scientists and engineers. Tables included here, for example, are those found more often in books on nonparametric statistics than in books likely to have been used in college courses for engineers. This book points the environmental and water resources scientist to robust and nonparametric statistics, and to exploratory data analysis. We believe that the characteristics of environmental (and perhaps most other 'real') data drive analysis methods towards use of robust and nonparametric methods.

Exercises are included at the end of chapters. In our course, students compute each type of analysis (t-test, regression, etc.) the first time by hand. We choose the smaller, simpler examples for hand computation. In this way the mechanics of the process are fully understood, and computer software is seen as less mysterious.

We wish to acknowledge and thank several other scientists at the U. S. Geological Survey for contributing ideas to this book. In particular, we thank those who have served as the other instructors at the USGS training course. Ed Gilroy has critiqued and improved much of the material found in this book. Tim Cohn has contributed in several areas, particularly to the sections on bias correction in regression, and methods for data below the reporting limit. Richard Alexander has added to the trend analysis chapter, and Charles Crawford has contributed ideas for regression and ANOVA. Their work has undoubtedly made its way into this book without adequate recognition.

Professor Ken Potter (University of Wisconsin) and Dr. Gary Tasker (USGS) reviewed the manuscript, spending long hours with no reward except the knowledge that they have improved the work of others. For that we are very grateful. We also thank Madeline Sabin, who carefully typed original drafts of the class notes on which the book is based. As always, the responsibility for all errors and slanted thinking are ours alone.

Dennis R. Helsel

Robert M. Hirsch

Reston, VA USA  
June, 1991